Advanced Regression: 1a Overview of the course

Garyfallos Konstantinoudis

Epidemiology and Biostatistics, Imperial College London

21 February 2023

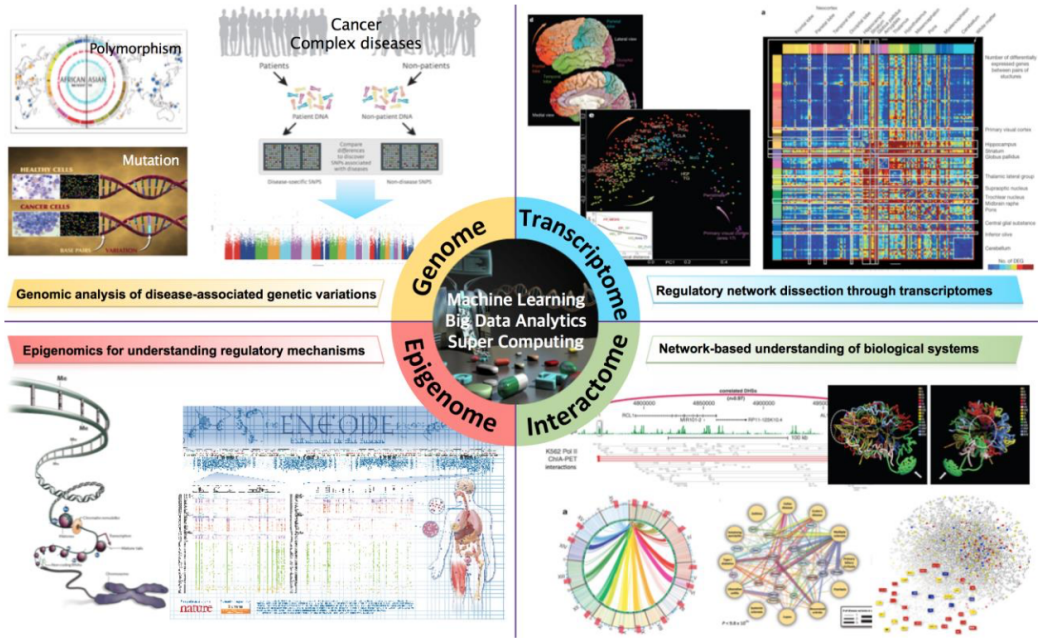
Advanced Regression: Motivation

- Course aims
- High-dimensional data

Advanced Regression: Course details

- Practicals
- Timetable
- Learning outcomes and exam
- Questions?

Advanced Regression: Course aims



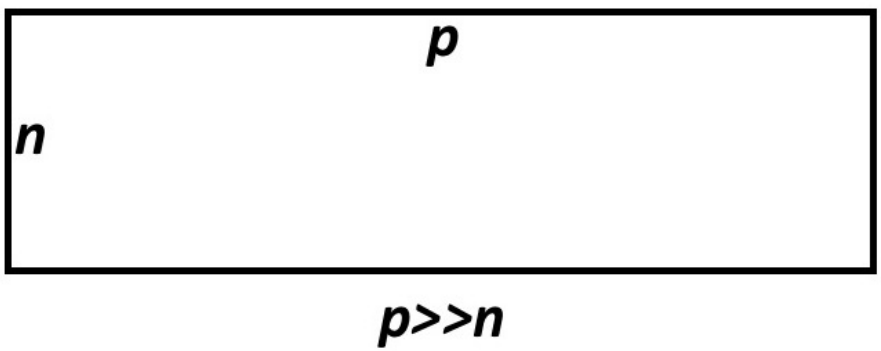
- Learn principles of advanced regression for high-dimensional data analysis.
- Apply these techniques on real-world data problems.

Motivation: High-dimensional data

- Number of samples or observations: n
- Number of variables: p

Data types:

- Big data: $p >> n$



- (Tall data: Summary-level data $p \times 1$)

Examples high-dimensional data types

- Health data records, e-records
- Health and fitness apps, location tracker
- Imaging data, e.g. functional and structural fMRI studies
- Credit scoring based on credit files and personal data
- Recommender systems based on user ratings

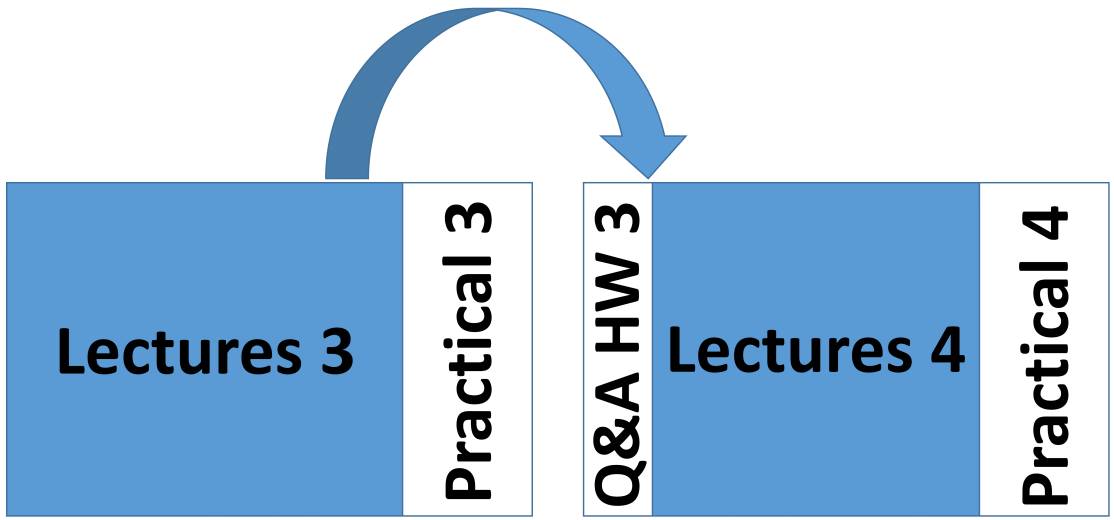
Modern data science is build on advanced regression models!

Methods covered in this course

- Random effects and hierarchical models
- Non-linear regression
- Penalised regression (Ridge, lasso, and elastic net)
- Building a prediction rule and cross-validation
- Classification with discriminant analysis and support vector machines
- Non-parametric methods (bagging, boosting, decision trees and random forests)
- Machine-learning models (Neural networks)

Advanced Regression: Practicals

- Structure: First lectures, then practicals and homework (optional) with one week delay



- Practical questions are available on blackboard. Solutions available online after the practical.
- Open discussion with module lead and tutors

Advanced Regression: Practicals

Each student needs to present the solution to one practical question.

- It is required to present one practical question to be admitted to the exam. (not this year)

Practicals are an essential part of the course.

- They help you to understand better the content and the relevance of the topics presented in the lectures.
- They are an important preparation for the exam.
- But more importantly, you will learn the basics of data science and statistical computing.

Why use R?

- The practicals will be in R.
- R is a language and environment for statistical computing and graphics.
- R is free and published under the GNU licence.
- 16,883 available add-on packages.
- Please download R and be prepared to run analysis before the practicals.
- Practical questions will be posted on Blackboard Thursday or Friday before the practical.
- `https://cran.r-project.org/`

What is markdown?

- When coding in R it is important to document and comment the code.
- Markdown is an R package that compiles R code into documents (pdf, html, word and many more).
- Package `https://cran.r-project.org/web/packages/rmarkdown`
- Project page `https://rmarkdown.rstudio.com/`

Make your code accessible and reproducible.

- Markdown can help you with that.
- Both practical questions and solutions will be provided in markdown and pdf format.

Week 1: 21st February

10:00-10:20	Lecture 1a	Overview and motivation	GK
10:20-11:00	Lecture 1b	Linear and generalised linear models	GK
11:10-12:00	Lecture 1c	Random effects and hierarchical models	GK
13:00-15:00	Practical 1	Using R to analyse data with linear models	GK & Christina

Week 2: 28th March

10:00-10:50	Lecture 2a	Introduction to non-linear regression	GK
11:00-11:50	Lecture 2b	Bias and variance trade off and penalised splines	GK
12:00-12:50	Lecture 2c	Distributed non-linear lag models	GK
14:00-16:00	Practical 2	Using R to perform non-linear regression	GK & CL

Week 3: 7th March

10:00-10:50	Lecture 2a	Variable selection	GK
11:00-11:50	Lecture 2b	Prediction accuracy and cross-validation	GK
12:00-12:50	Lecture 2c	Penalised regression models	GK
14:00-16:00	Practical 3	Using R to perform cross-validation and penalised regression	GK & CL

Week4: 14st March

10:00-10:50	Lecture 2a	Machine learning: Classification	GK
11:00-11:50	Lecture 2b	Machine learning: Ensemble methods	GK
12:00-12:50	Lecture 2c	Machine learning: Neural networks	GK
14:00-16:00	Practical 4	Using R to perform classification methods	GK & CL

Week 5: 21th March

10:00-12:00	Practical 5	Using R to understand ensemble methods	GK & CL
13:00-14:00	Mock exam	Go through the mock exam	GK
14:00-15:00	Q&A Session	Revisit concepts/lectures	GK

Advanced Regression: Learning outcomes

- Perform advanced statistical analyses, employing penalised likelihood or non-pararametric regression models.
- Discuss the theoretical foundations and limitations of the most widely used advanced regression approaches.
- Identify the challenges of high-dimensional data analysis.
- Identify suitable analysis strategies to address the problems arising from 'small n, large p' data sets.
- Use complex regression models in R, understand which methods are suitable for which data, know the pitfalls of high-dimensional data analysis, and interpret the results.
- Enjoy data science.

Advanced Regression: Exam

- This module will be assessed by a written open book programming exam which is a mixture of coding in R, interpretion of outputs, and description of why and how the analysis is performed.
- The exam is taking place in early May 2023 (tbc).
- Practical sessions offer regular opportunity for receiving formative feedback from the tutors.
- A mock exam will be provided and discussed.
- A Q & A session will be scheduled on the last day of the module and if helpful before the exam (End of April).

Advanced Regression: Questions?

Please get in touch:

g.konstantinoudis@imperial.ac.uk

Blackboard:

Discussion board, you can also post anonymously.

Drop-in sessions:

Every Wednesday 16:00-17:00 UK time

Zoom link will also be provided.

Next lectures

LECTURE 1b Linear models and generalised linear models

- Repetition: The linear model
- Generalised linear model

LECTURE 1c Random effects models

- Motivation: Structured data
- Fixed and random effects